Large Language Models Enable Few-Shot Clustering

Year: 2023 | Citations: 102 | Authors: Vijay Viswanathan, Kiril Gashteovski, Carolin (Haas) Lawrence

---

Abstract

Unlike traditional unsupervised clustering, semi-supervised clustering allows users to provide meaningful structure to the data, which helps the clustering algorithm to match the user's intent. Existing approaches to semi-supervised clustering require a significant amount of feedback from an expert to improve the clusters. In this paper, we ask whether a large language model (LLM) can amplify an expert's guidance to enable query-efficient, few-shot semi-supervised text clustering. We show that LLMs are surprisingly effective at improving clustering. We explore three stages where LLMs can be incorporated into clustering: before clustering (improving input features), during clustering (by providing constraints to the clusterer), and after clustering (using LLMs post-correction). We find that incorporating LLMs in the first two stages routinely provides significant improvements in cluster quality, and that LLMs enable a user to make trade-offs between cost and accuracy to produce desired clusters. We release our code and LLM prompts for the public to use.1